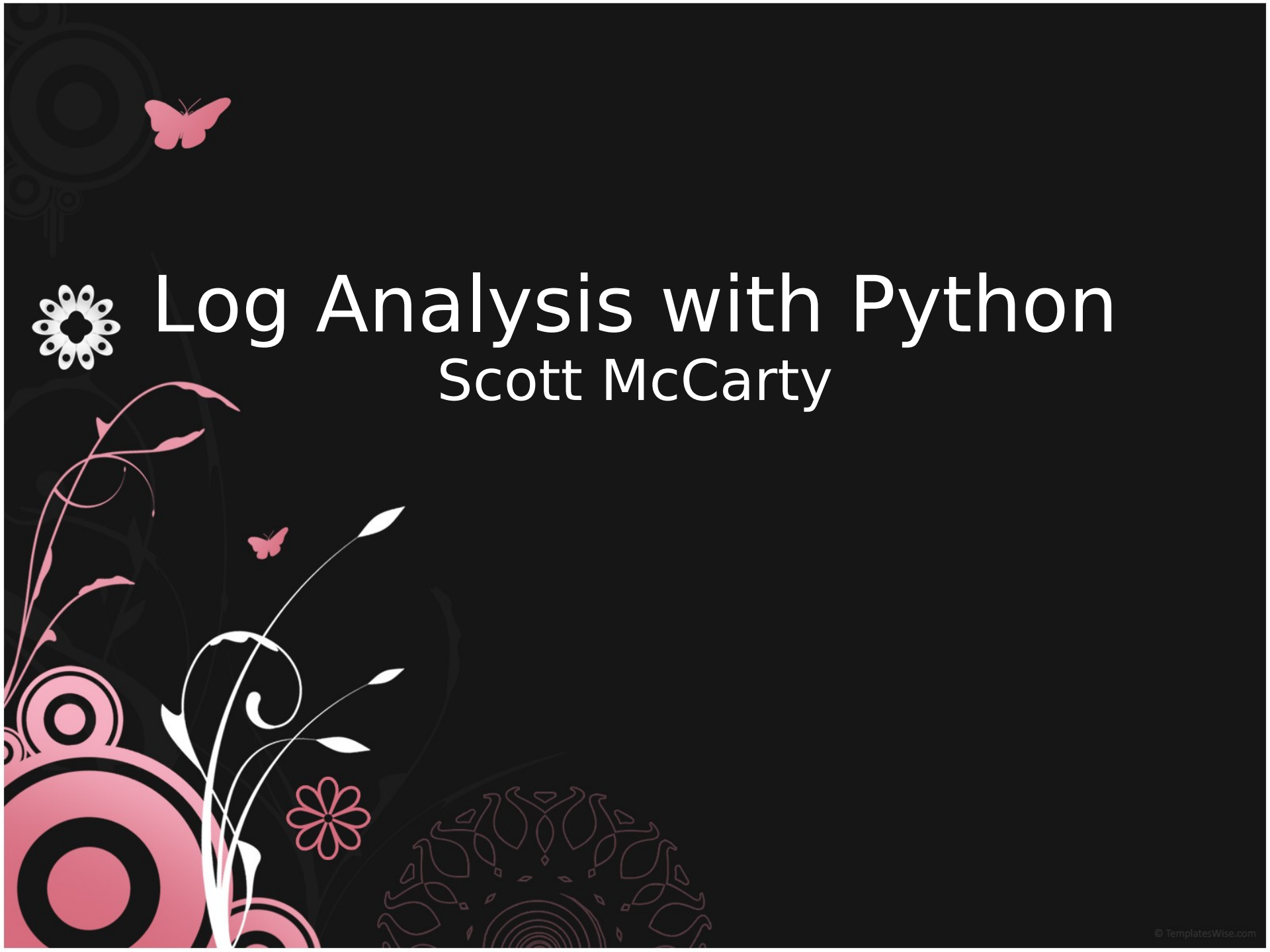
Log Analysis with Python

Scott McCarty



Why Log Analysis

- **Baselining**: Gathering statistics to find out what a normal time segment looks like
- **Reporting**: Proactive log analysis to give that warm and fuzzy feeling
- **Troubleshooting**: Quickly and efficiently tracking down a specifc problem in the logs

Why Python

- Culture & Community
- Attended college when CS was taught with C++ (easy transition)
- I was sick of PHP developers making fun of me using Perl (I'm not dead yet!)

Goals

- Provide standard quantitative and qualitative techniques in log analysis out of the box
- Provide easy to use command line tool
- Crunch various log types transparently
- Provide a library

Basics

- **Hashing**: Uses artificial ignorance to reduce the number of unique lines of text
- **Graphing**: Command line graphing
- **Wordcounts**: Anthropological technique to do word discovery
- **Other Counts**: Daemons, hosts, etc

Hashing

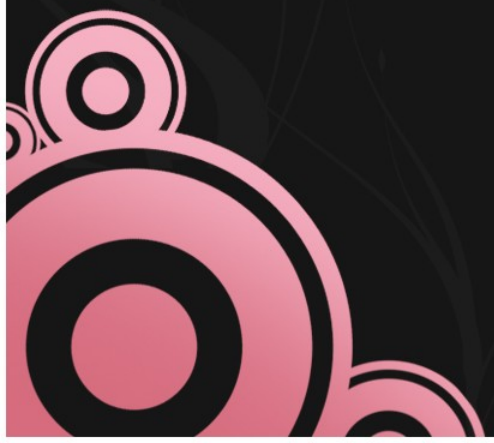
- We are looking for unique lines of text
- Remove numbers, Dates, IP Addresses, MAC addresses, etc
- This technique is sometimes referred to as artificial ignorance

Normal Syslog

```
[root@henry ~]# cat /root/test.log
Jul 28 23:55:01 henry snmpd[7774]: Connection from UDP: [10.0.8.52]:48853
Jul 28 23:55:01 henry snmpd[7774]: Received SNMP packet(s) from UDP: [10.0.8.52]
Jul 28 23:55:01 henry snmpd[7774]: Connection from UDP: [10.0.8.52]:48853
Jul 28 23:55:01 henry snmpd[7774]: Connection from UDP: [10.0.8.52]:48833
Jul 28 23:55:02 henry last message repeated 2 times
Jul 28 23:55:06 henry snmpd[7774]: Connection from UDP: [127.0.0.1]:36909
Jul 28 23:55:21 henry snmpd[7774]: Connection from UDP: [127.0.0.1]:36909
Jul 28 23:55:37 henry snmpd[7774]: Connection from UDP: [127.0.0.1]:36910
Jul 28 23:55:37 henry snmpd[7774]: Received SNMP packet(s) from UDP: [127.0.0.1]
Jul 28 23:55:52 henry snmpd[7774]: Connection from UDP: [127.0.0.1]:36910
[root@henry ~]#
```

Hashed Syslog

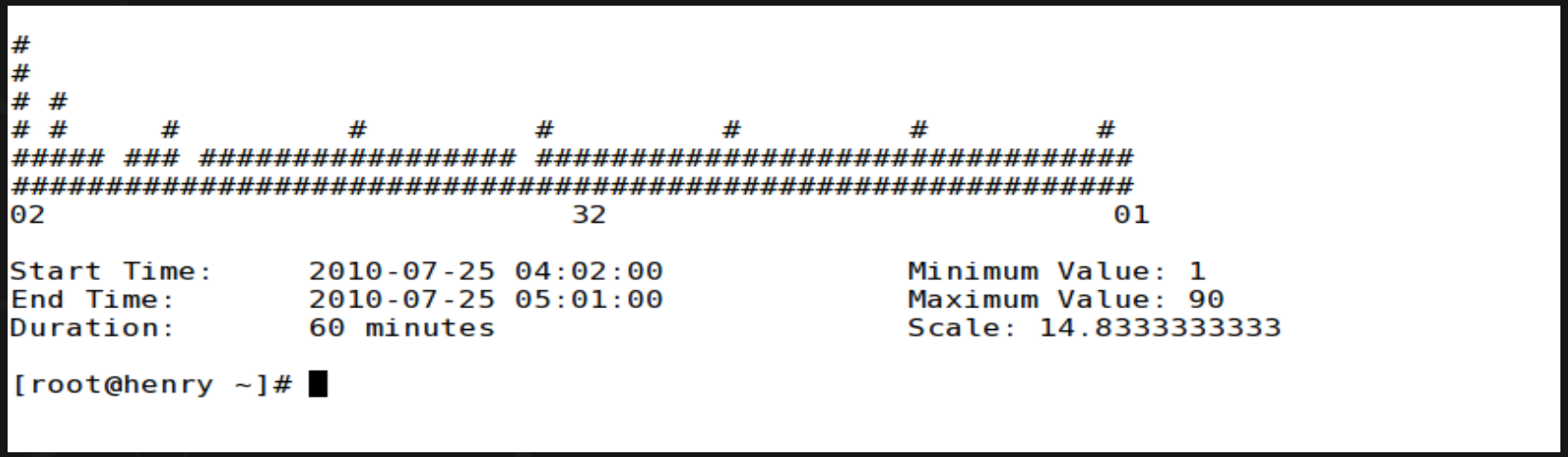
```
[root@henry ~]# cat /root/test.log | petit --hash
7:      snmpd[#]: Connection from UDP: [#.#.#.#]:#
2:      Received SNMP packet(s) from UDP: [10.0.8.52]
1:      message repeated 2 times
[root@henry ~]#
```



Graphing

- Simple command line graphs, used to track down problems
- Used to determine baseline/normal

Command Line Graphing



Wordcounts

- Natural language technique used in anthropology
- Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Vivamus et magna. Fusce sed sem sed magna suscipit egestas.

Wordcount Syslog

```
[root@henry ~]# tail /var/log/messages | petit --wordcount
5:      UDP:
4:      Connection
4:      message
4:      repeated
4:      times
3:      ##
1:      '#in-addr.arpa/IN'
1:      Received
1:      SNMP
1:      client
1:      denied
1:      internal:
1:      packet(s)
1:      update
1:      view
[root@henry ~]#
```

Other Counts

- **Hosts**: Counts number of messages per host
- **Daemons**: counts number of messages per daemon
- Good for baselining nad finding normal

Reporting

- Daily Reports: build an intimacy with the way things look normally, very qualitative in nature
- Monthly Reports: are a second chance to view the same data in the daily reports

Troubleshooting

- Quickly track down what went awry with hashing
- Quickly track down when something went wrong with graphing
- Use cat, grep, and petit to track down problems quickly

Future

- Standard deviation calculations
- Better fuzzy matches
- Better library